**Raport on lab2**

I selected Anime DataSet 2022 from Kaggle to perform my analysis. It has 13 columns and about 18k rows. After looking over data, I decided to use 'Type', 'Episodes', 'Studio', 'Tags', 'Release\_year' columns together with ‘Score’. All of them contain categorical data. Most interesting one were Tags. This column contained variety of tags, that anime has separated by a comma. It needed to be separated into different columns. I had to also deal with a lot of Nan’s, and do one hot encoding of all features. After that data was ready with over 1k columns and 15k rows. I decided to use DecisionTreeRegressor as my model. After training it had MSE error of 0.138.

I extracted 50 most important features from model, and a lot of them were expected. Tags like Drama, Sports, Sci-Fi or Action helped the model most while guessing rating. Model also favored anime from years 2016-2022, witch can be attributed to recency bias. Also unsurprisingly Some of most important features were well respected studios such as Studio Bones.

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznie

*Studio\_Unknown is a collective of all smaller studios indicating a bias towards big studios*

Lastly I Extracted all features that had 0 feature importance. It turned out to be quite a lot. From 1700 features, about 500 were not important at all to the model. Looking at them it makes sense. A lot of them are some obscure studios that only made 1 or 2 anime’s. Another large chunk were obscure tags that only a few anime’s had.

In conclusion, mosst feature importance’s were expected with a couple of interesting ones. Model heavily utilizes recency bias present in anime scores, and studio bias. It also managed to find some tags that resonate with people and other that people hated.

Filip Firkowski 151946